Gibbs Max-Margin Topic Models with Fast Sampling Algorithms

Jun Zhu DCSZJ@MAIL.TSINGHUA.EDU.CN
Ning Chen NINGCHEN@MAIL.TSINGHUA.EDU.CN
Hugh Perkins NGLS11@MAILS.TSINGHUA.EDU.CN
Bo Zhang DCSZB@MAIL.TSINGHUA.EDU.CN
Dept. of Comp. Sci & Tech; TNLIST Lab; State Key Lab of Intell. Tech & Sys., Beijing, 100084 China

Abstract

Existing max-margin supervised topic models rely on an iterative procedure to solve multiple latent SVM subproblems with additional mean-field assumptions on the desired posterior distributions. This paper presents Gibbs max-margin topic models by minimizing an expected margin loss, an upper bound of the existing margin loss derived from an expected prediction rule. By introducing augmented variables, we develop simple and fast Gibbs sampling algorithms with no restricting assumptions and no need to solve SVM subproblems for both classification and regression. Empirical results demonstrate significant improvements on time efficiency. The classification performance is also significantly improved over competitors.

1. Introduction

As supervising information gets easier to obtain, developing supervised latent topic models has attracted a lot of attentions. Both maximum likelihood estimation (MLE) and max-margin learning have been applied to learn supervised topic models. Different from the MLE-based approaches (Blei & McAuliffe, 2007), which define a normalized likelihood model for response variables, max-margin supervised topic models, such as maximum entropy discrimination LDA (MedLDA) (Zhu et al., 2009), directly minimize a margin-based loss derived from an expected prediction rule.

Although max-margin supervised topic models have shown superior performance in various settings, such as text mining (Zhu et al., 2009) and image annotation (Yang et al., 2010), their learning problems are generally hard to solve. Existing methods rely on a variational approximation scheme with strict mean-field assumptions on posterior distributions, and they normally need to solve multiple latent SVM subproblems in an EM-type iterative procedure. The recent work (Jiang et al., 2012) developed Monte Carlo (MC) methods for such max-margin topic models, with a weaker mean-field assumption; but they also need to solve multiple SVM problems. Thus, their efficiency could be limited as learning SVMs is normally computationally demanding. Also, it is not easy to parallelize these algorithms for large-scale applications.

This paper presents Gibbs MedLDA, a new formulation of max-margin supervised topic models with efficient inference algorithms. Instead of minimizing the margin loss of an expected prediction rule as adopted in MedLDA, Gibbs MedLDA minimizes the expected margin loss of many latent prediction rules, each rule corresponding to a configuration of topic assignments and the prediction model, drawn from a post-data posterior distribution. Theoretically, the expected margin loss is an upper bound of the existing margin loss of an expected prediction rule. Computationally, although the new margin loss can be hard in developing variational algorithms, we can develop simple and fast collapsed Gibbs sampling algorithms without any restricting assumptions on the posterior distribution, by exploiting the classical ideas of data augmentation (Dempster et al., 1977; Tanner & Wong, 1987; van Dyk & Meng, 2001) and its recent extensions to max-margin classifiers (Polson & Scott, 2011). We further generalize the ideas to develop a Gibbs MedLDA regression model and its Gibbs sampling algorithm with data augmentation. Empirical results on real data sets demonstrate significant improvements on time efficiency. The classification performance is also significantly improved.

The paper is organized as follows. Sec 2 reviews MedL-

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

DA and its EM-type algorithms. Sec 3 presents Gibbs MedLDA and its sampling algorithms for classification and regression. Sec 4 presents empirical results. Sec 5 concludes and discusses future directions.

2. MedLDA

We consider binary classification with a labeled training set $\mathcal{D} = \{(\mathbf{w}_d, y_d)\}_{d=1}^D$, where the response variable Y takes values from the output space $\mathcal{Y} = \{-1, +1\}$. MedLDA consists of two parts — an LDA model for describing input documents $\mathbf{W} = \{\mathbf{w}_d\}_{d=1}^D$, where $\mathbf{w}_d = \{w_{dn}\}_{n=1}^{N_d}$ denote the words appearing in document d, and an expected classifier for considering the supervising signal $\mathbf{y} = \{y_d\}_{d=1}^D$. Below, we introduce each of them in turn.

LDA: LDA is a hierarchical Bayesian model that posits each document as an admixture of K topics, where each topic $\boldsymbol{\Phi}_k$ is a multinomial distribution over a V-word vocabulary. For document d, the generating process can be described as

1. draw a topic proportion $\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\alpha})$
2. for each word n $(1 \le n \le N_d)$:
   (a) draw a topic assignment[1] $z_{dn} \sim \text{Mult}(\boldsymbol{\theta}_d)$
   (b) draw the observed word $w_{dn} \sim \text{Mult}(\boldsymbol{\Phi}_{z_{dn}})$

where $\text{Dir}(\cdot)$ is a Dirichlet distribution; $\text{Mult}(\cdot)$ is multinomial; and $\boldsymbol{\Phi}_{z_{dn}}$ denotes the topic selected by the non-zero entry of z_{dn}. The topics are random samples drawn from a prior, e.g., $\boldsymbol{\Phi}_k \sim \text{Dir}(\boldsymbol{\beta})$.

Given a set of documents $\mathbf{W}$, we let $\mathbf{z}_d = \{z_{dn}\}_{n=1}^{N_d}$, $\mathbf{Z} = \{\mathbf{z}_d\}_{d=1}^D$, and $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_d\}_{d=1}^D$. LDA infers the posterior distribution $p(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}|\mathbf{W}) \propto p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})$. We can show that the posterior distribution by Bayes' rule is the solution of an information theoretical optimization problem

$$\min_{q(\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})} \text{KL}[q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \| p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})] - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})]$$
$$\text{s.t.}: \; q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \in \mathcal{P}, \tag{1}$$

where $\text{KL}(q||p)$ is the Kullback-Leibler divergence, and $\mathcal{P}$ is the space of probability distributions. In fact, if we add the constant $\log p(\mathbf{W})$ to the objective, it is the minimization of $\text{KL}(q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \| p(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}|\mathbf{W}))$.

Expected Classifier: Given a training set $\mathcal{D}$, an expected classifier chooses a posterior distribution $q(h|\mathcal{D})$ over a hypothesis space $\mathcal{H}$ of classifiers such that the q-weighted (expected) classifier $h_q(\mathbf{w}) = \text{sign}\mathbb{E}_q[h(\mathbf{w})]$ will have the smallest possible risk. MedLDA follows this principle to learn a posterior $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}|\mathcal{D})$ such that the expected classifier

$$\hat{y} = \text{sign} F(\mathbf{w}) \tag{2}$$

has the smallest possible risk, approximated by the training error $\mathcal{R}_\mathcal{D}(q) = \sum_d \mathbb{I}(\hat{y}_d \neq y_d)$. The discriminant function is defined as

$$F(\mathbf{w}) = \mathbb{E}_{q(\boldsymbol{\eta},\mathbf{z}|\mathcal{D})}[F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w})], \; F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w}) = \boldsymbol{\eta}^\top \bar{\mathbf{z}}$$

where $\bar{\mathbf{z}}$ is a vector with element $\bar{z}_k = \frac{1}{N} \sum_{n=1}^N \mathbb{I}(z_n^k = 1)$, and $\mathbb{I}(\cdot)$ is an indicator function that equals to 1 if predicate holds otherwise 0. Note that the expected classifier and the LDA likelihood are coupled via the latent topic assignments $\mathbf{Z}$. The strong coupling makes it possible for MedLDA to learn a posterior distribution that can describe the observed words well and make accurate predictions.

Regularized Bayesian Inference: To integrate the above two components for hybrid learning, MedlDA regularizes the properties of the topic representations by imposing the following max-margin constraints derived from the classifier (2) to a standard LDA inference problem (1)

$$y_d F(\mathbf{w}_d) \ge \ell - \xi_d, \; \forall d, \tag{3}$$

where ℓ (≥ 1) is the cost of making a wrong prediction; and $\boldsymbol{\xi} = \{\xi_d\}_{d=1}^D$ are non-negative slack variables for inseparable cases. Let $\mathcal{L}(q) = \text{KL}(q||p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) - \mathbb{E}_q[\log p(\mathbf{W}|\mathbf{Z}, \boldsymbol{\Phi})]$ be the objective for doing standard Bayesian inference with the classifier $\boldsymbol{\eta}$ and $p_0(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) = p_0(\boldsymbol{\eta}) p_0(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$. MedLDA solves the regularized Bayesian inference (Zhu et al., 2011) problem

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) \in \mathcal{P}, \boldsymbol{\xi}} \mathcal{L}(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) + 2c \sum_d \xi_d \tag{4}$$
$$\forall d, \; \text{s.t.}: \; y_d F(\mathbf{w}_d) \ge \ell - \xi_d, \; \xi_d \ge 0,$$

where the margin constraints directly regularize the properties of the post-data distribution and c is the positive regularization parameter. Equivalently, MedLDA solves the unconstrained problem[2]

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})} \mathcal{L}(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) + 2c\mathcal{R}(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) \tag{5}$$

where $\mathcal{R}(q) = \sum_d \max(0, \ell - y_d F(\mathbf{w}_d))$ is the hinge loss that upper bounds the training error $\mathcal{R}_\mathcal{D}(q)$ of the expected classifier (2). Note that the factor 2 is included simply for convenience.

2.1. Existing Iterative Algorithms

Since it is intractable to solve problem (4) or (5) directly, both variational and Monte Carlo methods have been developed for approximate solutions. It can be shown that the variational method (Zhu et al., 2012) is a coordinate descent algorithm to solve problem (5) with the fully-factorized assumption that

[1] A K-dim binary vector with only one nonzero entry.

[2] If not specified, q is subject to the constraint $q \in \mathcal{P}$.

$q(\boldsymbol{\eta}, \Theta, \mathbf{Z}, \boldsymbol{\Phi}) = q(\boldsymbol{\eta})(\prod_d q(\boldsymbol{\theta}_d) \prod_n q(z_{dn})) \prod_k q(\boldsymbol{\Phi}_k)$, while the Monte Carlo methods (Jiang et al., 2012) make a weaker assumption that $q(\boldsymbol{\eta}, \Theta, \mathbf{Z}, \boldsymbol{\Phi}) = q(\boldsymbol{\eta}) q(\Theta, \mathbf{Z}, \boldsymbol{\Phi})$. All these methods have a similar EM-type iterative procedure, which solves many latent SVM subproblems, as outlined below.

Estimate $q(\boldsymbol{\eta})$: Given $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$, this step solves

$$\begin{aligned} &\min_{q(\boldsymbol{\eta}), \boldsymbol{\xi}} \quad \mathrm{KL}(q(\boldsymbol{\eta}) \| p_0(\boldsymbol{\eta})) + 2c \sum_d \xi_d \\ &\forall d, \text{ s.t.}: \quad y_d \mathbb{E}_q[\boldsymbol{\eta}]^\top \mathbb{E}_q[\bar{\mathbf{z}}_d] \geq \ell - \xi_d, \ \xi_d \geq 0. \end{aligned} \tag{6}$$

When the prior $p_0(\boldsymbol{\eta})$ is the commonly used standard normal, we have the optimum solution $q(\boldsymbol{\eta}) = \mathcal{N}(\boldsymbol{\kappa}, I)$, where $\boldsymbol{\kappa} = \sum_d y_d \mu_d \mathbb{E}_q[\bar{\mathbf{z}}_d]$ and μ_d are Lagrange multipliers. It can be shown that the dual problem of (6) is the dual of a standard binary linear SVM and we can solve it or its primal form efficiently using existing high-performance SVM learners. We denote the optimum solution of this problem by $(q^*(\boldsymbol{\eta}), \boldsymbol{\kappa}^*, \boldsymbol{\xi}^*, \boldsymbol{\mu}^*)$.

Estimate $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$: Given $q(\boldsymbol{\eta})$, this step solves

$$\begin{aligned} &\min_{q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}), \boldsymbol{\xi}} \ \mathcal{L}(q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) + 2c \sum_d \xi_d \\ &\forall d, \text{ s.t.}: y_d (\boldsymbol{\kappa}^*)^\top \mathbb{E}_q[\bar{\mathbf{z}}_d] \geq \ell - \xi_d, \ \xi_d \geq 0. \end{aligned} \tag{7}$$

Although we can solve this problem using Lagrangian methods, it would be hard to derive the dual objective. An effective approximation strategy was used in (Zhu et al., 2012; Jiang et al., 2012), which updates $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ for only one step with $\boldsymbol{\xi}$ fixed at $\boldsymbol{\xi}^*$. By fixing $\boldsymbol{\xi}$ at $\boldsymbol{\xi}^*$, we have the solution $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \propto p(\mathbf{W}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi}) \exp\{(\boldsymbol{\kappa}^*)^\top \sum_d \mu_d^* \bar{\mathbf{z}}_d\}$, where the second term indicates the regularization effects due to the max-margin posterior constraints. For those data with non-zero Lagrange multipliers (i.e., support vectors), the second term will bias MedLDA towards a new posterior distribution that favors more discriminative representations on these "hard" data points. The Monte Carlo methods directly draw samples from the posterior distribution $q(\boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ or its collapsed form using Gibbs sampling to estimate $\mathbb{E}_q[\bar{\mathbf{z}}_d]$, the expectations required to learn $q(\boldsymbol{\eta})$. In contrast, the variational methods solve problem (7) using coordinate descent to estimate $\mathbb{E}_q[\bar{\mathbf{z}}_d]$ with a fully factorized assumption.

3. Gibbs MedLDA

Now, we present Gibbs max-margin topic models and their "augment-and-collapse" sampling algorithms.

3.1. Learning with an Expected Margin Loss

As stated above, MedLDA chooses the strategy to minimize the hinge loss of an expected classifier. In learning theory, an alternative approach to building classifiers with a posterior distribution of models is to minimize an expected loss, under the framework known as Gibbs classifiers (or stochastic classifiers) (McAllester, 2003; Catoni, 2007; Germain et al., 2009) with nice theoretical properties.

For our case of inferring the distribution of latent topic assignments $\mathbf{Z}$ and the classification model $\boldsymbol{\eta}$, the expected margin loss is defined as follows. If we have drawn a sample of the topic assignments $\mathbf{Z}$ and the prediction model $\boldsymbol{\eta}$ from a posterior distribution $q(\boldsymbol{\eta}, \mathbf{Z})$, we can define the linear discriminant function $F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w}) = \boldsymbol{\eta}^\top \bar{\mathbf{z}}$ as before and make prediction using the latent Gibbs rule

$$\hat{y} = \mathrm{sign} F(\boldsymbol{\eta}, \mathbf{z}; \mathbf{w}). \tag{8}$$

Let $\zeta_d = \ell - y_d \boldsymbol{\eta}^\top \bar{\mathbf{z}}_d$. The hinge loss of the classifier is $\mathcal{R}(\boldsymbol{\eta}, \mathbf{Z}) = \sum_d \max(0, \zeta_d)$ and the expected hinge loss is

$$\mathcal{R}'(q) = \mathbb{E}_q[\mathcal{R}(\boldsymbol{\eta}, \mathbf{Z})] = \sum_d \mathbb{E}_q[\max(0, \zeta_d)].$$

Since $\mathcal{R}(\boldsymbol{\eta}, \mathbf{Z}) \geq \sum_d \mathbb{I}(y_d \neq \hat{y}_d)$ for any $(\boldsymbol{\eta}, \mathbf{Z})$, we have $\mathcal{R}'(q) \geq \sum_d \mathbb{E}_p[\mathbb{I}(y_d \neq \hat{y}_d)]$. In other words, $\mathcal{R}'(q)$ is an upper bound of the expected training error of the Gibbs classifier (8). Thus, it is a good surrogate loss for training.

Then, with the same goal as MedLDA to find a posterior distribution $q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})$ that on one hand describes the observed data and on the other hand predicts as well as possible on training data, we define Gibbs MedLDA as solving the new regularized Bayesian inference problem

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathcal{L}(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) + 2c \mathcal{R}'(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})). \tag{9}$$

Comparing to MedLDA in problem (5), we have the following lemma by applying Jensen's inequality.

Lemma 1 *The expected hinge loss $\mathcal{R}'$ is an upper bound of the hinge loss of the expected classifier (2):*

$$\mathcal{R}'(q) \geq \sum_d \max(0, \mathbb{E}_q[\zeta_d]).$$

3.2. Formulation with Data Augmentation

If we directly solve problem (9), the expected hinge loss $\mathcal{R}'$ is hard to deal with because we do not have a closed-form of the expectation of the max function. Fortunately, we can develop a simple collapsed Gibbs sampling method based on a data augmentation formulation of the expected hinge-loss.

Let $\phi(y_d | \mathbf{z}_d, \boldsymbol{\eta}) = \exp\{-2c \max(0, \zeta_d)\}$ be the unnormalized pseudo-likelihood of the response variable for document d. Then, problem (9) can be written as

$$\min_{q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})} \mathcal{L}(q(\boldsymbol{\eta}, \boldsymbol{\Theta}, \mathbf{Z}, \boldsymbol{\Phi})) - \mathbb{E}_q[\log \phi(\mathbf{y} | \mathbf{Z}, \boldsymbol{\eta})], \tag{10}$$

where $\phi(\mathbf{y}|\mathbf{Z},\boldsymbol{\eta}) = \prod_d \phi(y_d|\boldsymbol{\eta},\mathbf{z}_d)$. Solving problem (10) with the constraint that $q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) \in \mathcal{P}$, we can get the normalized posterior distribution

$$q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = \frac{p_0(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})\phi(\mathbf{y}|\mathbf{Z},\boldsymbol{\eta})}{\psi(\mathbf{y},\mathbf{W})},$$

where $\psi(\mathbf{y},\mathbf{W})$ is the normalization constant. Using the ideas of data augmentation (Tanner & Wong, 1987; Polson & Scott, 2011), we have Lemma 2.

Lemma 2 (Scale of Mixture) *The unnormalized pseudo-likelihood can be expressed as*

$$\phi(y_d|\mathbf{z}_d,\boldsymbol{\eta}) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_d}} \exp\big(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\big) d\lambda_d$$

Proof: Due to the fact that $a\max(0,x) = \max(0,ax)$ if $a \geq 0$, we have $-2c\max(0,\zeta_d) = -2\max(0,c\zeta_d)$. Then, we can follow the proof in (Polson & Scott, 2011) to get the results. □

Lemma 2 indicates that the posterior distribution of Gibbs MedLDA, $q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$, can be expressed as the marginal of a higher dimensional distribution that includes the augmented variables $\boldsymbol{\lambda}$. The complete posterior distribution is

$$q(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) = \frac{p_0(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})p(\mathbf{W}|\mathbf{Z},\boldsymbol{\Phi})\phi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},\boldsymbol{\eta})}{\psi(\mathbf{y},\mathbf{W})},$$

where the pseudo-joint distribution of $\mathbf{y}$ and $\boldsymbol{\lambda}$ is

$$\phi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},\boldsymbol{\eta}) = \prod_d \frac{1}{\sqrt{2\pi\lambda_d}} \exp\big(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\big).$$

In fact, we can show that the complete posterior distribution is the solution of the data augmentation problem of Gibbs MedLDA

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})} \mathcal{L}(q(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})) - \mathbb{E}_q[\log p(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},\boldsymbol{\eta})],$$

which is again subject to the normalization constraint that $q(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}) \in \mathcal{P}$.

3.3. Inference with Collapsed Gibbs Sampling

Although we can do Gibbs sampling to infer the complete posterior distribution $q(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$ and thus $q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})$ by ignoring $\boldsymbol{\lambda}$, the mixing rate would be slow due to the large sample space. One way to effectively reduce the sample space and improve mixing rates is to integrate out the intermediate Dirichlet variables $(\boldsymbol{\Theta},\boldsymbol{\Phi})$ and build a Markov chain whose equilibrium distribution is the resulting marginal distribution $q(\boldsymbol{\eta},\boldsymbol{\lambda},\mathbf{Z})$. We propose to use collapsed Gibbs sampling, which has been successfully used in LDA (Griffiths & Steyvers, 2004). With the data augmentation representation, this leads to an "augment-and-collapse" sampling algorithm for Gibbs MedLDA.

For Gibbs MedLDA, the collapsed posterior distribution is

$$\begin{aligned} q(\boldsymbol{\eta},\boldsymbol{\lambda},\mathbf{Z}) &\propto p_0(\boldsymbol{\eta})p(\mathbf{W},\mathbf{Z}|\boldsymbol{\alpha},\boldsymbol{\beta})\phi(\mathbf{y},\boldsymbol{\lambda}|\mathbf{Z},\boldsymbol{\eta}) \\ &= p_0(\boldsymbol{\eta})\Big[\prod_{d=1}^D \frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})}\Big]\prod_{k=1}^K \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})} \\ &\quad \times \prod_{d=1}^D \frac{1}{\sqrt{2\pi\lambda_d}} \exp\Big(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\Big), \end{aligned}$$

where $\delta(\mathbf{x}) = \frac{\prod_{i=1}^{\dim(\mathbf{x})}\Gamma(x_i)}{\Gamma(\sum_{i=1}^{\dim(\mathbf{x})} x_i)}$, C_k^t is the number of times the term t being assigned to topic k over the whole corpus and $\mathbf{C}_k = \{C_k^t\}_{t=1}^V$; C_d^k is the number of times that terms being associated with topic k within the d-th document and $\mathbf{C}_d = \{C_d^k\}_{k=1}^K$. Then, the conditional distributions used in collapsed Gibbs sampling are as follows.

For $\boldsymbol{\eta}$: let's assume its prior is an isotropic Gaussian distribution $p_0(\boldsymbol{\eta}) = \prod_k \mathcal{N}(\eta_k; 0, \nu^2)$. Then, we have

$$\begin{aligned} q(\boldsymbol{\eta}|\mathbf{Z},\boldsymbol{\lambda}) &\propto p_0(\boldsymbol{\eta})\prod_d \exp\big(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\big) \\ &\propto \exp\big(-\sum_k \frac{\eta_k^2}{2\nu^2} - \sum_d \frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\big) \\ &= \mathcal{N}(\boldsymbol{\eta};\boldsymbol{\mu},\Sigma), \end{aligned} \tag{11}$$

where the posterior mean is $\boldsymbol{\mu} = \Sigma(c\sum_d y_d \frac{\lambda_d + c\ell}{\lambda_d}\bar{\mathbf{z}}_d)$ and the covariance matrix is $\Sigma = (\frac{1}{\nu^2}I + c^2\sum_d \frac{\bar{\mathbf{z}}_d\bar{\mathbf{z}}_d^\top}{\lambda_d})^{-1}$. Therefore, we can easily draw a sample from a K-dimensional multivariate Gaussian distribution. The inverse can be robustly done using Cholesky decomposition, an $O(K^3)$ procedure. Since K is normally not large, the inversion can be done efficiently.

For $\mathbf{Z}$: The conditional distribution of $\mathbf{Z}$ is

$$\begin{aligned} q(\mathbf{Z}|\boldsymbol{\eta},\boldsymbol{\lambda}) &\propto \prod_{d=1}^D \frac{\delta(\mathbf{C}_d + \boldsymbol{\alpha})}{\delta(\boldsymbol{\alpha})} \exp\big(-\frac{(\lambda_d + c\zeta_d)^2}{2\lambda_d}\big) \\ &\quad \times \prod_{k=1}^K \frac{\delta(\mathbf{C}_k + \boldsymbol{\beta})}{\delta(\boldsymbol{\beta})}. \end{aligned}$$

By canceling common factors, we can derive the conditional distribution of one variable z_{dn} given others $\mathbf{Z}_\neg$ as:

$$\begin{aligned} &q(z_{dn}^k = 1|\mathbf{Z}_\neg,\boldsymbol{\eta},\boldsymbol{\lambda},w_{dn} = t) \\ &\propto \frac{(C_{k,\neg n}^t + \beta_t)(C_{d,\neg n}^k + \alpha_k)}{\sum_t C_{k,\neg n}^t + \sum_{t=1}^V \beta_t} \exp\Big(\frac{\gamma y_d(c\ell + \lambda_d)\eta_k}{\lambda_d} \\ &\quad - c^2\frac{\gamma^2\eta_k^2 + 2\gamma(1-\gamma)\eta_k\Lambda_{dn}^k}{2\lambda_d}\Big), \end{aligned} \tag{12}$$

where $C_{\cdot,\neg n}^{\cdot}$ indicates that term n is excluded from the corresponding document or topic; $\gamma = \frac{1}{N_d}$; and $\Lambda_{dn}^k = \frac{1}{N_d - 1}\sum_{k'}\eta_{k'}C_{d,\neg n}^{k'}$ is the discriminant function value without word n. We can see that the first term

is from the LDA model for observed word counts and the second term is from the supervised signal $\mathbf{y}$.

For $\boldsymbol{\lambda}$: Finally, the conditional distribution of the augmented variables $\boldsymbol{\lambda}$ is

$$\begin{aligned} q(\lambda_d|\mathbf{Z},\boldsymbol{\eta}) &\propto \frac{1}{\sqrt{2\pi\lambda_d}}\exp\big(-\frac{(\lambda_d+c\zeta_d)^2}{2\lambda_d}\big) \\ &\propto \frac{1}{\sqrt{2\pi\lambda_d}}\exp\big(-\frac{c^2\zeta_d^2}{2\lambda_d}-\frac{\lambda_d}{2}\big) \\ &= \mathcal{GIG}\big(\lambda_d;\frac{1}{2},1,c^2\zeta_d^2\big), \end{aligned}$$

where $\mathcal{GIG}(x;p,a,b)=C(p,a,b)x^{p-1}\exp(-\frac{1}{2}(\frac{b}{x}+ax))$ is a generalized inverse Gaussian distribution (Devroye, 1986) and $C(p,a,b)$ is a normalization constant. Therefore, we can derive that λ_d^{-1} follows an inverse Gaussian distribution

$$p(\lambda_d^{-1}|\mathbf{Z},\boldsymbol{\eta})=\mathcal{IG}\big(\lambda_d^{-1};\frac{1}{c|\zeta_d|},1\big), \tag{13}$$

where $\mathcal{IG}(x;a,b)=\sqrt{\frac{b}{2\pi x^3}}\exp(-\frac{b(x-a)^2}{2a^2x})$ for $a>0$ and $b>0$.

With the above conditional distributions, we can construct a Markov chain which iteratively draws samples of $\boldsymbol{\eta}$ using Eq. (11), $\mathbf{Z}$ using Eq. (12) and $\boldsymbol{\lambda}$ using Eq. (13), with an initial condition. To sample from an inverse Gaussian distribution, we apply the transformation method with multiple roots (Michael et al., 1976). In our experiments, we initially set $\boldsymbol{\lambda}=1$ and randomly draw $\mathbf{Z}$ from a uniform distribution. In training, we run this Markov chain to finish the burn-in stage with M iterations. Then, we draw a sample $\hat{\boldsymbol{\eta}}$ as the Gibbs classifier to make predictions on testing data.

3.4. Prediction

To apply the Gibbs classifier $\hat{\boldsymbol{\eta}}$, we need to infer the topic assignments for testing document. We take the approach in (Zhu et al., 2012; Jiang et al., 2012), which uses a point estimate of topics $\boldsymbol{\Phi}$ from training data and makes prediction based on them. Specifically, we use the MAP estimate $\hat{\boldsymbol{\Phi}}$ to replace the probability distribution $p(\boldsymbol{\Phi})$. For the collapsed Gibbs sampler, an estimate of $\hat{\boldsymbol{\Phi}}$ using the samples is $\hat{\phi}_{kt}\propto C_k^t+\beta_t$. Then, given a testing document $\mathbf{w}$, we infer its latent components $\mathbf{z}$ using $\hat{\boldsymbol{\Phi}}$ as $p(z_n^k=1|\mathbf{z}_{\neg n})\propto\hat{\phi}_{kw_n}(C_{\neg n}^k+\alpha_k)$, where $C_{\neg n}^k$ is the times that the terms in this document $\mathbf{w}$ assigned to topic k with the n-th term excluded.

3.5. Gibbs MedLDA Regression Model

Before ending this section, we briefly discuss how to generalize the above ideas to develop a regression model, where the response variable Y takes real values. Specifically, the Gibbs MedLDA regression model has the same LDA model to describe input words and a Gibbs regression model for the response variable. If a sample of the topic assignments $\mathbf{Z}$ and the prediction model $\boldsymbol{\eta}$ is drawn from the posterior distribution $q(\boldsymbol{\eta},\mathbf{Z})$, we define the prediction rule as $\hat{y}=\boldsymbol{\eta}^\top\mathbf{z}$. One widely used margin-based loss measure is the ϵ-insensitive loss $\mathcal{R}_\epsilon(\boldsymbol{\eta},\mathbf{Z})=\sum_d\max(0,|\Delta_d|-\epsilon)$ for support vector regression (Smola & Scholkopf, 2003), where $\Delta_d=y_d-\boldsymbol{\eta}^\top\bar{\mathbf{z}}_d$ is the margin. Then, we define the Gibbs MedLDA regression model as solving the regularized Bayesian inference problem

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})}\mathcal{L}(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}))+2c\mathcal{R}_\epsilon(q(\boldsymbol{\eta},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})), \tag{14}$$

where $\mathcal{R}_\epsilon=\mathbb{E}_q[\mathcal{R}_\epsilon(\boldsymbol{\eta},\mathbf{Z})]=\sum_d\mathbb{E}_q[\max(0,|\Delta_d|-\epsilon)]$ is the expected ϵ-insensitive loss. Similarly, we can show that $\mathcal{R}_\epsilon$ is an upper bound of the ϵ-insensitive loss of MedLDA's expected prediction rule, by noting that

$$\max(0,|x|-\epsilon)=\max(0,x-\epsilon)+\max(0,-x-\epsilon). \tag{15}$$

Lemma 3 *We have $\mathcal{R}_\epsilon\geq\sum_d\max(0,|\mathbb{E}_q[\Delta_d]|-\epsilon)$.*

Proof: By using the equality (15), we have $\mathcal{R}_\epsilon=\sum_d(\mathbb{E}_q[\max(0,\Delta_d-\epsilon)]+\mathbb{E}_q[\max(0,-\Delta_d-\epsilon)])$. Analogous to Lemma 1, we can show that $\mathbb{E}_q[\max(0,\Delta_d-\epsilon)]\geq\max(0,\mathbb{E}_q[\Delta_d]-\epsilon)$ and $\mathbb{E}_q[\max(0,-\Delta_d-\epsilon)]\geq\max(0,-\mathbb{E}_q[\Delta_d]-\epsilon)$. Then, applying the equality (15) again, we get the results. □

We can reformulate problem (14) in the form as problem (10), with the pseudo-likelihood $\phi(y_d|\boldsymbol{\eta},\mathbf{z}_d)=\exp(-2c\max(0,|\Delta_d|-\epsilon))$. Then, we have the dual scale of mixture representation.

Lemma 4 (Dual Scale of Mixture) *For regression, the pseudo-likelihood can be expressed as*

$$\begin{aligned}\phi(y_d|\boldsymbol{\eta},\mathbf{z}_d)=&\int_0^\infty\frac{1}{\sqrt{2\pi\lambda_d}}\exp\big(-\frac{(\lambda_d+c(\Delta_d-\epsilon))^2}{2\lambda_d}\big)d\lambda_d\\ &\times\int_0^\infty\frac{1}{\sqrt{2\pi\omega_d}}\exp\big(-\frac{(\omega_d-c(\Delta_d+\epsilon))^2}{2\omega_d}\big)d\omega_d\end{aligned}$$

Proof: By the equality (15), we have $\phi(y_d|\boldsymbol{\eta},\mathbf{z}_d)=\exp\{-2c\max(0,\Delta_d-\epsilon)\}\exp\{-2c\max(0,-\Delta_d-\epsilon)\}$. Each of the exponential terms can be formulated as a scale mixture of Gaussians due to Lemma 2. □

Then, the data augmented learning problem of the Gibbs MedLDA regression model is

$$\min_{q(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi})}\mathcal{L}(q(\boldsymbol{\eta},\boldsymbol{\lambda},\boldsymbol{\omega},\boldsymbol{\Theta},\mathbf{Z},\boldsymbol{\Phi}))-\mathbb{E}_q[\log\phi(\mathbf{y},\boldsymbol{\lambda},\boldsymbol{\omega}|\mathbf{Z},\boldsymbol{\eta})]$$

where $\phi(\mathbf{y},\boldsymbol{\lambda},\boldsymbol{\omega}|\mathbf{Z},\boldsymbol{\eta})=\prod_d\phi(y_d,\lambda_d,\omega_d|\mathbf{Z},\boldsymbol{\omega})$ and

$$\begin{aligned}\phi(y_d,\lambda_d,\omega_d|\mathbf{Z},\boldsymbol{\eta})=&\frac{1}{\sqrt{2\pi\lambda_d}}\exp\big(-\frac{(\lambda_d+c(\Delta_d-\epsilon))^2}{2\lambda_d}\big)\\ &\times\frac{1}{\sqrt{2\pi\omega_d}}\exp\big(-\frac{(\omega_d-c(\Delta_d+\epsilon))^2}{2\omega_d}\big).\end{aligned}$$

Solving the augmented problem and integrating out $(\boldsymbol{\Theta}, \boldsymbol{\Phi})$, we can get the collapsed posterior distribution

$$q(\boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{Z}) \propto p_0(\boldsymbol{\eta}) p(\mathbf{W}, \mathbf{Z}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \phi(\mathbf{y}, \boldsymbol{\lambda}, \boldsymbol{\omega}|\mathbf{Z}, \boldsymbol{\eta}).$$

Then, following similar derivations as in the classification model, the Gibbs sampling algorithm to infer the posterior has the following conditional distributions.

For $\boldsymbol{\eta}$: again, with the isotropic Gaussian prior $p_0(\boldsymbol{\eta}) = \prod_k \mathcal{N}(\eta_k; 0, \nu^2)$, we have

$$q(\boldsymbol{\eta}|\mathbf{Z}, \boldsymbol{\lambda}, \boldsymbol{\omega}) = \mathcal{N}(\boldsymbol{\eta}; \boldsymbol{\mu}, \Sigma), \tag{16}$$

where $\Sigma = (\frac{1}{\nu^2} I + c^2 \sum_d \rho_d \bar{\mathbf{z}}_d \bar{\mathbf{z}}_d^\top)^{-1}$, $\boldsymbol{\mu} = c\Sigma(\sum_d \psi_d \bar{\mathbf{z}}_d)$, $\rho_d = \frac{1}{\lambda_d} + \frac{1}{\omega_d}$ and $\psi_d = \frac{y_d - \epsilon}{\lambda_d} + \frac{y_d + \epsilon}{\omega_d}$. We can easily draw a sample from a K-dimensional multivariate Gaussian distribution. The inverse can be robustly done using Cholesky decomposition.

For $\mathbf{Z}$: We can derive the conditional distribution of one variable z_{dn} given others $\mathbf{Z}_\neg$ as:

$$\begin{aligned} & q(z_{dn}^k = 1|\mathbf{Z}_\neg, \boldsymbol{\eta}, \boldsymbol{\lambda}, \boldsymbol{\omega}, w_{dn} = t) \\ & \propto \frac{(C_{k,\neg n}^t + \beta_t)(C_{d,\neg n}^k + \alpha_k)}{\sum_t C_{k,\neg n}^t + \sum_{t=1}^V \beta_t} \exp\Big(c\gamma\psi_d\eta_k \\ & - c^2\big(\frac{\gamma^2\zeta_d\eta_k^2}{2} + \gamma(1-\gamma)\rho_d\eta_k\Upsilon_{dn}^k\big)\Big), \end{aligned} \tag{17}$$

where $\gamma = \frac{1}{N_d}$; and $\Upsilon_{dn}^k = \frac{1}{N_d - 1}\sum_{k'} \eta_{k'} C_{d,\neg n}^{k'}$ is the discriminant function value without word n. The first term is from the LDA model for observed word counts. The second term is from the supervised signal $\mathbf{y}$.

For $\boldsymbol{\lambda}$: Finally, we can derive that λ_d^{-1} and ω_d^{-1} follow the inverse Gaussian distributions:

$$q(\lambda_d^{-1}|\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\omega}) = \mathcal{IG}\big(\lambda_d^{-1}; \frac{1}{c|\Delta_d - \epsilon|}, 1\big), \tag{18}$$

$$q(\omega_d^{-1}|\mathbf{Z}, \boldsymbol{\eta}, \boldsymbol{\lambda}) = \mathcal{IG}\big(\omega_d^{-1}; \frac{1}{c|\Delta_d + \epsilon|}, 1\big). \tag{19}$$

4. Experiments

We present empirical results to demonstrate the efficiency and prediction performance of Gibbs MedLDA (denoted by GibbsMedLDA) on the 20Newsgroups data set for classification and a hotel review data set for regression. We also analyze its sensitivity to key parameters. The 20Newsgroups data set contains about 20K postings within 20 groups. We follow the same setting as in (Zhu et al., 2012) and remove a standard list of stop words for both binary and multi-class classification. For all the experiments, we use the standard normal prior $p_0(\boldsymbol{\eta})$ (i.e., $\nu^2 = 1$) and the symmetric Dirichlet priors $\boldsymbol{\alpha} = \frac{\alpha}{K}\mathbf{1}$, $\boldsymbol{\beta} = 0.01 \times \mathbf{1}$, where $\mathbf{1}$ is a vector with all entries being 1. For each setting, we report the average performance and standard deviation with five randomly initialized runs. All the experiments are done on a standard desktop computer.

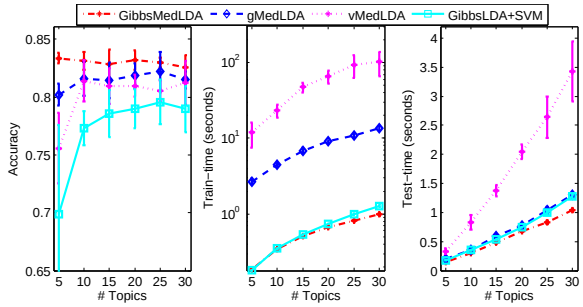

Figure 1. Classification accuracy, training time (in log-scale) and testing time (in linear scale) on the 20Newsgroups binary classification data set.

4.1. Binary classification

The binary classification is to distinguish postings of the newsgroup *alt.atheism* and postings of the group *talk.religion.misc*. The training set contains 856 documents, and the test set contains 569 documents. We compare Gibbs MedLDA with the MedLDA model that uses variational methods (denoted by vMedLDA) (Zhu et al., 2012) and the MedLDA that uses collapsed Gibbs sampling algorithms (denoted by gMedLDA) (Jiang et al., 2012). We also include the unsupervised LDA using collapsed Gibbs sampling as a baseline, denoted by GibbsLDA. For GibbsLDA, we learn a binary linear SVM on its topic representations using SVMLight (Joachims, 1999). The results of other supervised topic models, such as sLDA and DiscLDA (Lacoste-Jullien et al., 2009), were reported in (Zhu et al., 2012). For Gibbs MedLDA, we set $\alpha = 1$, $\ell = 164$ and $M = 10$. As we shall see, Gibbs MedLDA is insensitive to α, ℓ and M in a wide range. Although tuning c (e.g., via cross-validation) can produce slightly better results, we fix $c = 1$ for simplicity.

Fig. 1 shows the accuracy, training time, and testing time of different methods with different numbers of topics. We can see that by minimizing an expected hinge-loss and not making any restricting assumptions on the posterior distributions, GibbsMedLDA achieves higher accuracy than other max-margin topic models, which make some restricting assumptions. Similarly, as gMedLDA makes a weaker mean-field assumption, it achieves slightly higher accuracy than vMedLDA, which assumes that the posterior distribution is fully factorized. For the training time, GibbsMedLDA is about two orders of magnitudes faster than vMedLDA, and about one order of magnitude faster than gMedLDA. This is partly because both vMedLDA and gMedLDA need to solve multiple SVM problems. For the testing time, GibbsMedLDA is comparable with gMedLDA and the unsupervised GibbsLDA, but much faster than the variational algorithm used by vMedLDA, especially when K is large.

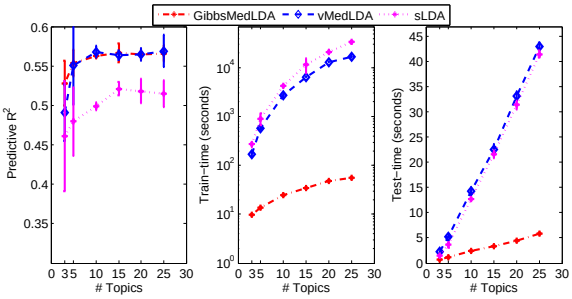

Figure 2. Predictive R2, training time and testing time on the hotel review data set.

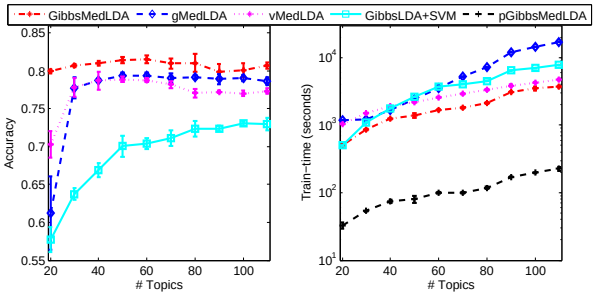

Figure 3. Classification accuracy and training time on the whole 20Newsgroups data set.

4.2. Regression

We use the hotel review data set (Zhu & Xing, 2010) built by randomly crawling hotel reviews from the TripAdvisor website where each review is associated with a global rating score ranging from 1 to 5. In these experiments, we focus on predicting the global rating scores for reviews using the bag-of-words features only, with a vocabulary of 12,000 terms. All the reviews have character lengths between 1500 and 6000. The data set consists of 5,000 reviews, with 1000 reviews per rating. The data set is uniformly partitioned into training and testing sets. We compare the Gibbs MedLDA regression model with the MedLDA regression model that uses variational inference and supervised LDA (sLDA) which also uses variational inference. For Gibbs MedLDA and vMedLDA, the precision is set at $\epsilon = 1e^{-3}$ and c is selected via 5 fold cross-validation during training. Again, we set the Dirichlet parameter $\alpha = 1$ and the number of burn-in $M = 10$.

Fig. 2 shows the predictive R^2 (Blei & McAuliffe, 2007) of different methods. We can see that GibbsMedLDA achieves comparable prediction performance with vMedLDA, which is better than sLDA. Note that vMedLDA uses a full likelihood model for both words and response variables, while GibbsMedLDA uses a simpler likelihood model for words only. For train time, GibbsMedLDA is about two orders of magnitudes faster than vMedLDA (as well as sLDA), again due to the fact that GibbsMedLDA doesn't need to solve multiple SVM problems. For testing time, GibbsMedLDA is also much faster than vMedLDA and sLDA, especially when the number of topics is large.

4.3. More discussions

4.3.1. Multi-class classification

We perform multi-class classification on the 20Newsgroups data with all 20 categories. The test set consists of 7,505 documents, and the training set consists of 11,269 documents. Again, since GibbsMedLDA is insensitive to α and ℓ, we set $\alpha = 1$ and $\ell = 64$. We also fix $c = 1$ for simplicity. The number of burn-in iterations is set as $M = 20$, which is sufficiently large, as we shall see.

Various methods exist to apply binary classifiers to do multi-class classification, including the popular "one-vs-all" and "one-vs-one" strategies. Here we choose the "one-vs-all" strategy, which has been shown effective (Rifkin & Klautau, 2004), to provide some preliminary analysis. Fig. 3 shows the classification accuracy and training time, where GibbsMedLDA builds 20 binary GibbsMedLDA classifiers. Since there is no coupling among these 20 binary classifiers, we can learn them in parallel, which we denote by pGibbsMedLDA. We can see a clear improvement on the classification accuracy, which may be due to the different strategies on building the multi-class classifiers[3]. However, given the performance gain on the binary classification task, we believe that the Gibbs sampling algorithm without any restricting factorization assumptions is another factor leading to the improved performance. For training time, GibbsMedLDA takes slightly less time than the variational MedLDA as well as gMedLDA. But if we train the multiple classifiers in parallel, we can save a lot of training time. These results are promising since it is now not uncommon to have a desktop computer with multiple processors or a cluster with tens or hundreds of computing nodes.

4.3.2. Sensitivity analysis

Burn-In: Fig. 4 shows the performance of GibbsMedLDA with different numbers of burn-in samples for the binary classification task. When $M = 0$, the model is in fact random. We can see that the classification performance increases very fast and converges to the stable optimum with 5 to 10 burn-in steps. The training time increases about linearly in general when using more burn-in steps. Moreover, the training time increases linearly as K increases. In the previous experiments, we have chosen $M = 10$.

[3]MedLDA learns multi-class SVM (Zhu et al., 2012).

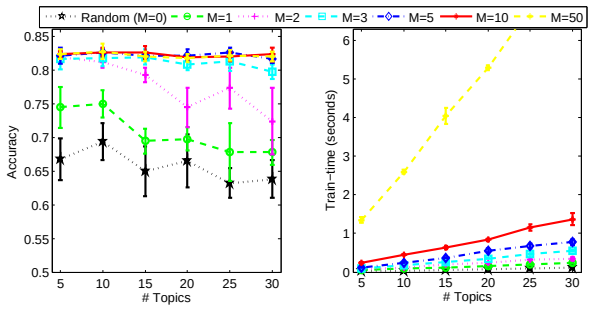

Figure 4. (L) accuracy and (R) training time of GibbsMedLDA with different numbers of burn-in steps for binary classification.

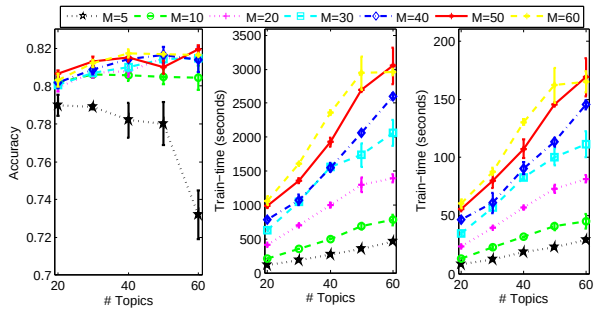

Figure 5. (L) accuracy, (M) training time of GibbsMedLDA and (R) training time of pGibbsMedLDA with different numbers of burn-in steps for multi-class classification.

Fig. 5 shows the performance of GibbsMedLDA for multi-class classification with different numbers of burn-in steps. We can see when the number of burn-in steps is larger than 20, the performance is quite stable. Again, the training time grows about linearly as the number of burn-in steps increases. Even if we use 40 or 60 steps of burn-in, the training time is still competitive, compared with the variational MedLDA, especially considering that GibbsMedLDA can be naively parallelized by learning different binary classifiers simultaneously.

Dirichlet prior α: Fig. 6 shows the classification performance of GibbsMedLDA on the binary task with different α values. For the three different topic numbers, we can see that the performance is quite stable in a wide range of α values, e.g., from 0.1 to 10. We can also see that it generally needs a larger α in order to get the best results when K becomes larger. This is mainly because a large K tends to produce sparse topic representations and an appropriately large α is needed to smooth the representations, as the effective Dirichlet prior is $\alpha_k = \alpha/K$.

Loss penalty ℓ: Fig. 7 shows the classification performance of GibbsMedLDA on the binary classification task with different ℓ values. Again, we can see that in a wide range, e.g., from 25 to 625, the performance is quite stable for all the three different K values. In

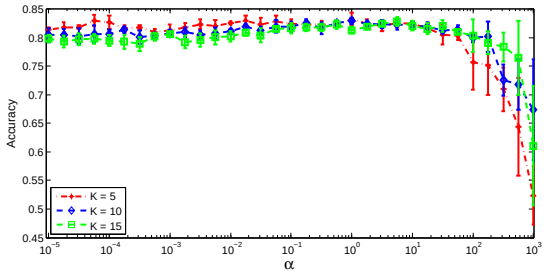

Figure 6. Classification accuracy of GibbsMedLDA on the binary classification data set with different α values.

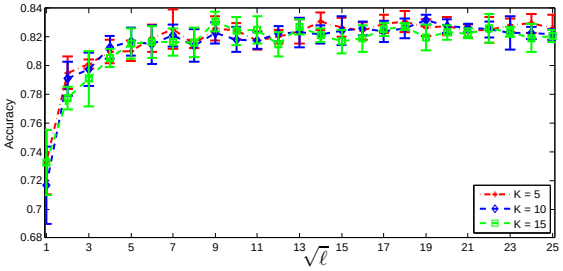

Figure 7. Classification accuracy of GibbsMedLDA on the binary classification data set with different ℓ values.

the above experiments, we set $\ell = 164$. For the multi-class classification task, we have similar observations, and we set $\ell = 64$ in the previous experiments.

5. Conclusions and Discussions

We presented Gibbs MedLDA, a new formulation of max-margin supervised topic models, which minimizes an expected margin loss. By using the idea of data augmentation, we presented simple and highly efficient "augment-and-collapse" Gibbs sampling algorithms without making any restricting assumptions on posterior distributions. Empirical results on real data demonstrate significant improvements over the existing max-margin topic models.

The new data augmentation formulation without any need to solve constrained subproblems has shown great promise on improving the time efficiency of max-margin topic models. For future work, we are interested in developing highly scalable sampling algorithms (e.g., using a distributed architecture) (Newman et al., 2009; Smola & Narayanamurthy, 2010) to deal with large scale data sets.

Acknowledgments

This work is supported by National Key Foundation R&D Projects (No.s 2013CB329403, 2012CB316301), Tsinghua Initiative Scientific Research Program No.20121088071, the 221 Basic Research Plan for Young Faculties at Tsinghua University, and a Research Fund No. 20123000007 from Microsoft Research Asia.

References

Blei, D.M. and McAuliffe, J.D. Supervised topic models. *Advances in Neural Information Processing Systems (NIPS)*, pp. 121–128, 2007.

Catoni, O. PAC-Bayesian supervised classification: The thermodynamics of statistical learning. *Monograph series of the Institute of Mathematical Statistics*, 2007.

Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, (39):1–38, 1977.

Devroye, L. *Non-uniform random variate generation*. Springer-Verlag, 1986.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *International Conference on Machine Learning (ICML)*, pp. 353–360, 2009.

Griffiths, T.L. and Steyvers, M. Finding scientific topics. *Proceedings of National Academy of Science (PNAS)*, pp. 5228–5235, 2004.

Jiang, Q., Zhu, J., Sun, M., and Xing, E.P. Monte Carlo methods for maximum margin supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Joachims, T. *Making large-scale SVM learning practical*. MIT press, 1999.

Lacoste-Jullien, S., Sha, F., and Jordan, M.I. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems (NIPS)*, pp. 897–904, 2009.

McAllester, D. PAC-Bayesian stochastic model selection. *Machine Learning*, 51:5–21, 2003.

Michael, J.R., Schucany, W.R., and Haas, R.W. Generating random variates using transformations with multiple roots. *The American Statistician*, 30(2): 88–90, 1976.

Newman, D., Asuncion, A., Smyth, P., and Welling, M. Distributed algorithms for topic models. *Journal of Machine Learning Research (JMLR)*, (10):1801–1828, 2009.

Polson, N.G. and Scott, S.L. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1): 1–24, 2011.

Rifkin, R. and Klautau, A. In defense of one-vs-all classification. *Journal of Machine Learning Research (JMLR)*, (5):101–141, 2004.

Smola, A. and Narayanamurthy, S. An architecture for parallel topic models. *Very Large Data Base (VLDB)*, 3(1-2):703–710, 2010.

Smola, A. and Scholkopf, B. A tutorial on support vector regression. *Statistics and Computing*, 14(3): 199–222, 2003.

Tanner, M.A. and Wong, W.-H. The calculation of posterior distributions by data augmentation. *Journal of the Americal Statistical Association (JASA)*, 82(398):528–540, 1987.

van Dyk, D. and Meng, X. The art of data augmentation. *Journal of Computational and Graphical Statistics (JCGS)*, 10(1):1–50, 2001.

Yang, S., Bian, J., and Zha, H. Hybrid generative/discriminative learning for automatic image annotation. In *Uncertainty in Artificial Intelligence (UAI)*, 2010.

Zhu, J. and Xing, E.P. Conditional topic random fields. In *International Conference on Machine Learning (ICML)*, pp. 1239–1246, 2010.

Zhu, J., Ahmed, A., and Xing, E.P. MedLDA: maximum margin supervised topic models for regression and classification. In *International Conference on Machine Learning (ICML)*, pp. 1257–1264, 2009.

Zhu, J., Chen, N., and Xing, E.P. Infinite latent SVM for classification and multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1620–1628, 2011.

Zhu, J., Ahmed, A., and Xing, E.P. MedLDA: maximum margin supervised topic models. *Journal of Machine Learning Research (JMLR)*, (13):2237–2278, 2012.